

UNIWIN VERSION 10.2.0

CLASSIFICATION PAR LA METHODE DES K-MEDOIDES

Révision : 25/03/2025

Définition.....	1
Entrée des données	2
Données manquantes	3
Exemple 1 : Fichier PAM	3
L'option Rapports	6
L'option Graphiques	9
Exemple 2 : Fichier VEHICULE	10
Exemple 3 : Fichier IRIS3.....	17
Exemple 4 : Fichier IRIS - Classification mixte	21
Les variables internes créées par la procédure	25
Références	25

Définition

La classification par la méthode des K-médoïdes est une approche de classification apparentée à la méthode des K-moyennes pour partitionner un ensemble de données en k classes. Dans la classification par K-médoïdes, chaque classe est représentée par l'une des observations de la classe. Ces observations sont appelées médoïdes.

Le terme médoïde fait référence à une observation au sein d'une classe pour laquelle la dissemblance moyenne entre elle et toutes les autres observations de la classe est minimale. Elle correspond au point le plus central de la classe. Ces observations (une par classe) peuvent être considérées comme des exemples représentatifs des membres des classes. Rappelons que, dans la classification par K-moyennes, le centre d'une classe donnée est calculé comme la valeur moyenne de toutes les observations de cette classe.

La méthode des K-médoïdes est une alternative robuste à la méthode des K-moyennes. Cela signifie que l'algorithme est moins sensible au bruit et aux valeurs aberrantes car il utilise les médoïdes comme centres des classes au lieu des moyennes.

La méthode des K-médoïdes la plus courante est l'algorithme PAM (Partitioning Around Medoids) de Kaufman et Rousseeuw (1990).

La procédure affiche un rapport indiquant notamment les médoïdes des classes formées, la classification des observations, des statistiques descriptives pour les classes formées et les contributions des variables aux classes. Si une classification mixte a été mise en œuvre, les résultats de la CAH sont également fournis.

Les graphiques des coefficients moyens et individuels de silhouette et des nuages de points des classes formées sont proposés. Si une classification mixte a été mise en œuvre, le diagramme des indices de la classification et l'affichage de l'arbre sont proposés.

Cette procédure est basée sur les packages R 'stats' et 'cluster'.

Entrée des données

Cliquons sur l'icône KM dans le ruban Décrire et choisissons K-médoïdes pour afficher la boîte de dialogue montrée ci-dessous :

Classification par la méthode des K-médoïdes

Variables quantitatives :

(Libellés des observations :)

(Libellés des variables :)

Nombre désiré de classes ou lignes des médoïdes initiaux :

Nombre maximum de classes à tester : 10

Nombre de tirages aléatoires : 10

Racine aléatoire : 1890383467

Faire une CAH sur les classes obtenues par PAM

Nombre désiré de classes pour la CAH : 1

Standardisation des données

Ok Annuler Sélection Supprimer Aide

Cette boîte de dialogue permet de choisir les variables quantitatives à utiliser pour la classification, la variable contenant les libellés des observations et la variable contenant les libellés des variables quantitatives utilisées pour la classification.

Par défaut les données sont standardisées par soustraction de la moyenne et division par l'écart absolu moyen (MAD).

Le nombre désiré de classes peut être précisé de deux façons, soit en entrant ce nombre, soit en entrant les numéros des lignes du fichier des données définissant les médoïdes initiaux de ces classes.

Le champ 'Nombre maximum de classes à tester' permet de calculer diverses statistiques pouvant aider à valider le nombre adéquat de classes à former.

Le nombre maximum d'itérations de l'algorithme peut être précisé.

Le nombre de tirages aléatoires des médoïdes de départ et la racine aléatoire associée peuvent être indiqués dans le cas où le nombre de classes à former est précisé. Ces deux options ne sont pas utilisées si les médoïdes initiaux des classes à former sont entrés.

Enfin, il est possible de compléter l'analyse par une classification ascendante hiérarchique (CAH) sur les classes obtenues par les K-médoïdes.

Données manquantes

Les données manquantes ne sont pas autorisées par cette procédure.

Exemple 1 : Fichier PAM

Pour ce premier exemple, nous utiliserons le fichier PAM.

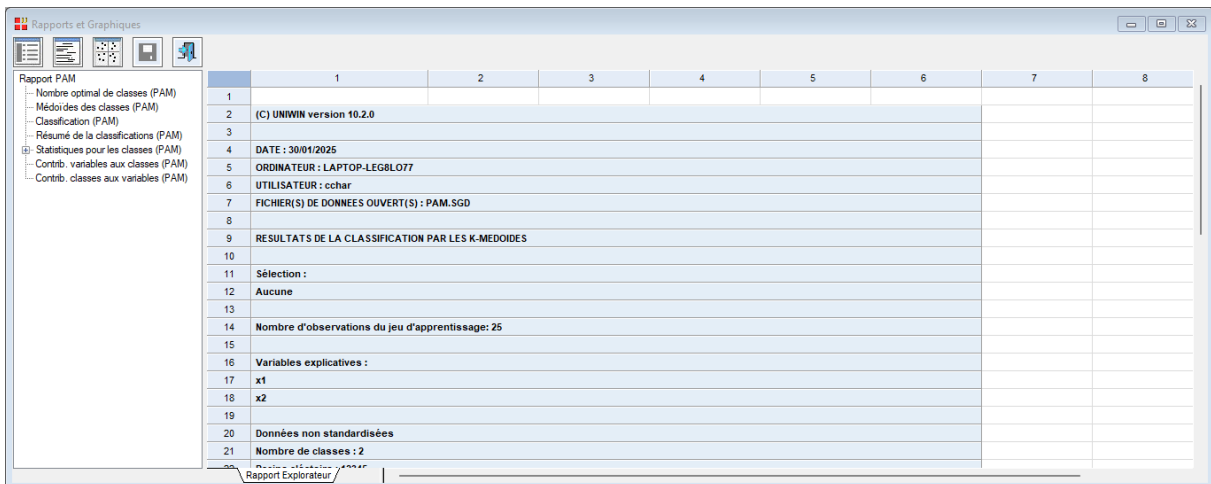
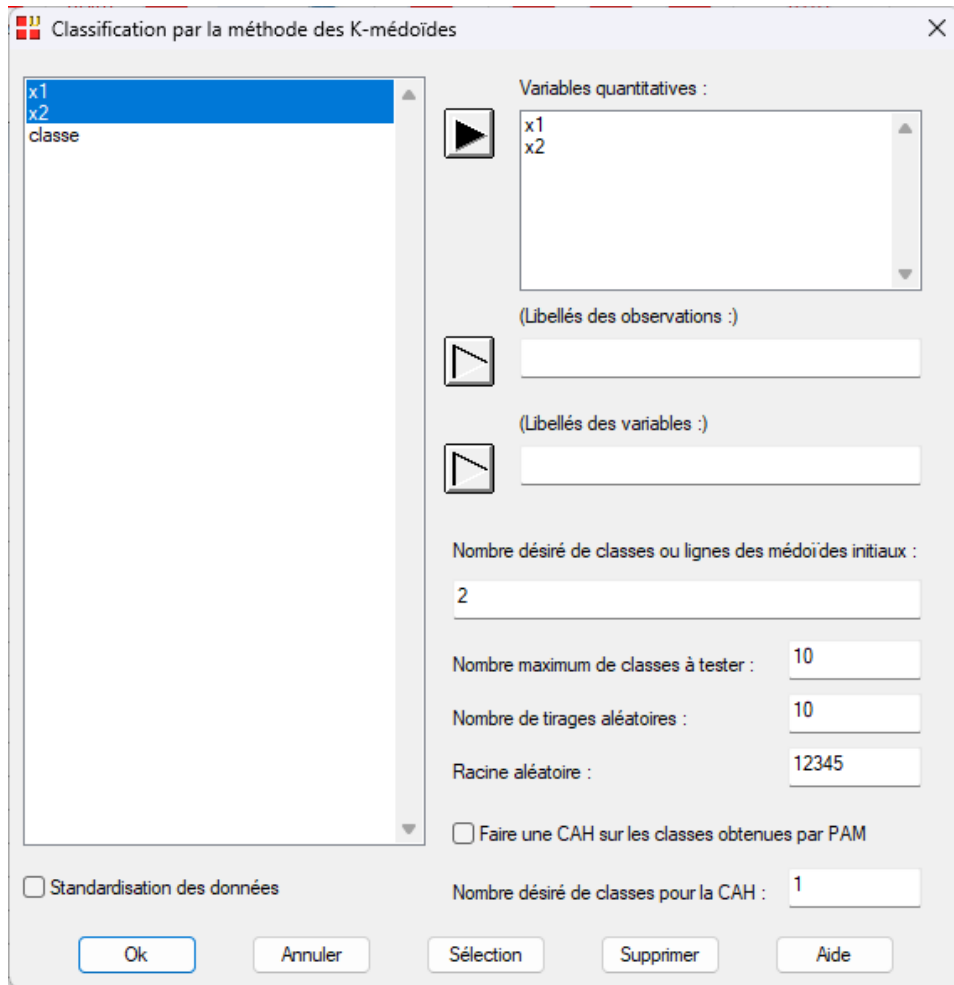
Il contient 25 observations (10 pour la classe 1 et 15 pour la classe 2) et deux variables x1 et x2 dont les données sont issues de tirages aléatoire de lois normales.


Cliquons sur l'icône KM dans le ruban Décrire et choisissons K-médoïdes et renseignons la boîte de dialogue montrée ci-après.


Nous demandons 2 classes et de faire une classification sur les données non standardisées.

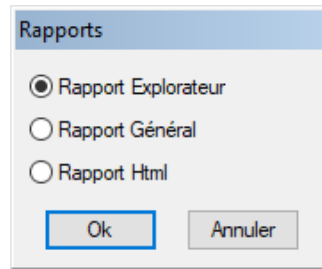
Cliquons sur Ok. UNIWIN débute le calcul de la classification.


Après quelques instants, l'écran suivant s'affiche :

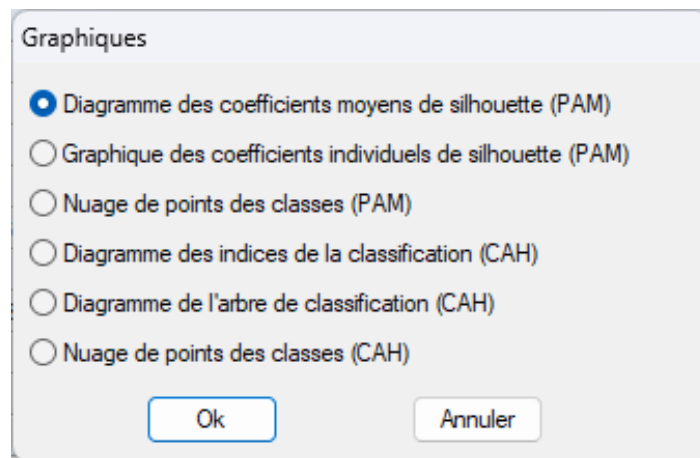


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

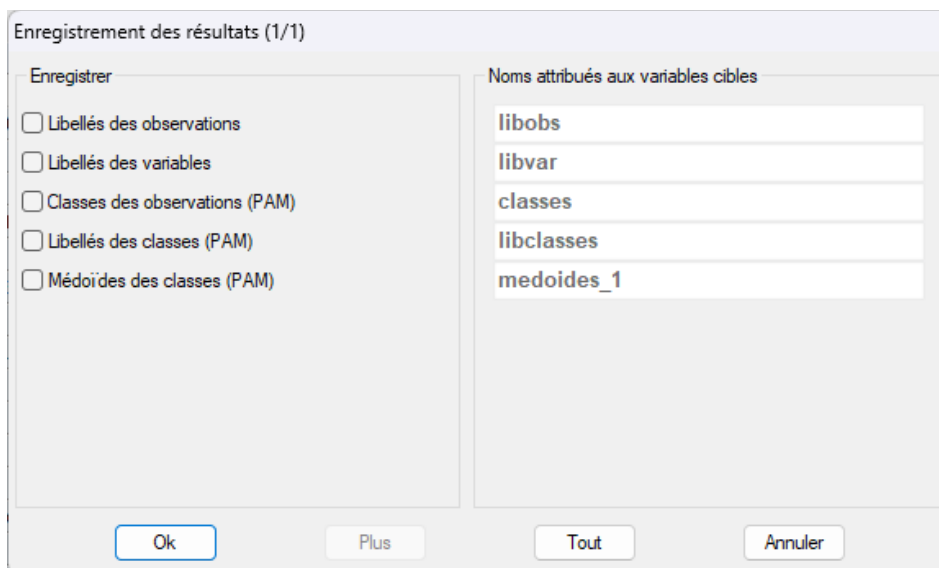
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



et l'icône 'Graphiques'  affiche la boîte de dialogue des options pour les graphiques.



L'icône 'Enregistrer'  permet de sélectionner les résultats de l'analyse à enregistrer dans un fichier.

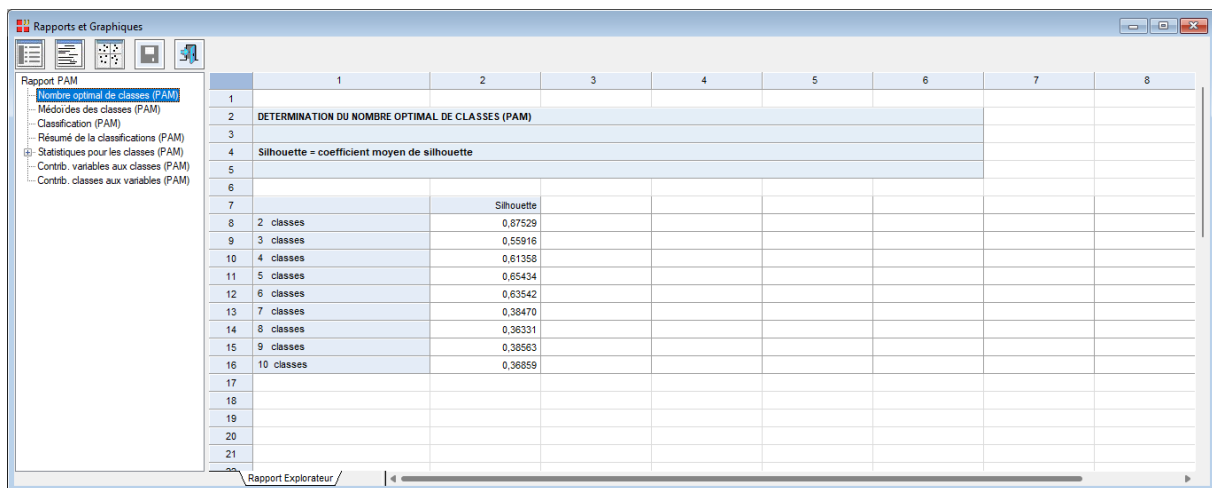


L'option Rapports

Cette option permet d'obtenir le rapport à l'écran sous la forme d'un tableur ou au format HTML.

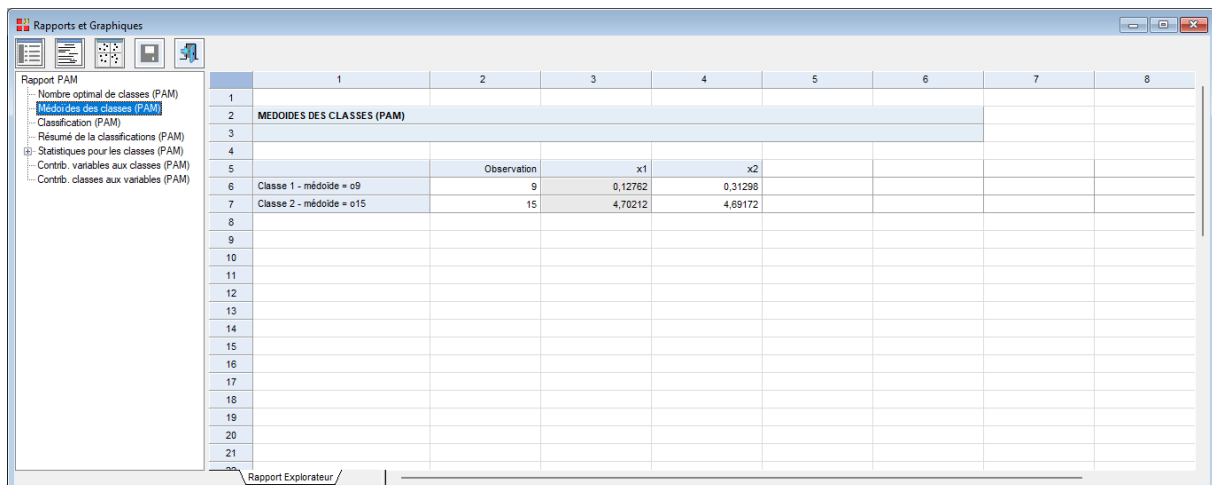
Le premier tableau affiche pour les nombres de classes variant de 2 au nombre maximum de classes indiqué (10), les coefficients moyens de silhouette.

Ces coefficients sont également représentés de façon graphique. Ils aident à déterminer le nombre adéquat de classes à former.



	1	2	3	4	5	6	7	8
1								
2	DETERMINATION DU NOMBRE OPTIMAL DE CLASSES (PAM)							
3								
4	Silhouette = coefficient moyen de silhouette							
5								
6								
7								
8	2 classes							
9	3 classes							
10	4 classes							
11	5 classes							
12	6 classes							
13	7 classes							
14	8 classes							
15	9 classes							
16	10 classes							
17								
18								
19								
20								
21								

Le deuxième tableau affiche les médoïdes des classes formées.



	1	2	3	4	5	6	7	8
1								
2	MEDOÏDES DES CLASSES (PAM)							
3								
4								
5								
6	Classe 1 - médoïde = o9							
7	Classe 2 - médoïde = o15							
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

La classe 1 est représentée par l'observation 9 et la classe 2 par l'observation 15.

Le troisième tableau affiche pour chaque observation sa classe d'affectation, sa classe voisine, sa distance au médoïde et son coefficient de silhouette.

	1	2	3	4	5	6	7	8
1								
2	CLASSIFICATION DES OBSERVATIONS (PAM)							
3								
4	Nombre de classes formées : 2							
5								
6	Distance = distance de l'observation au médoïde de sa classe							
7	Silhouette = coefficient de silhouette de l'observation							
8								
9								
		Classe affectée	Classe voisine	Distance	Silhouette			
11	o1	1	2	0,69307	0,88774			
12	o2	1	2	0,16766	0,93188			
13	o3	1	2	0,41425	0,91273			
14	o4	1	2	0,57925	0,90111			
15	o5	1	2	0,54746	0,88420			
16	o6	1	2	0,40033	0,91210			
17	o7	1	2	0,37422	0,90444			
18	o8	1	2	0,76699	0,85011			
19	o9	1	2	0,00000	0,93331			
20	o10	1	2	0,05580	0,93134			
21	o11	2	1	0,64388	0,87345			

Le quatrième tableau affiche un résumé de la classification.

	1	2	3	4	5	6	7	8
1								
2	INFORMATIONS SUR LES CLASSES (PAM)							
3								
4	Classe isolée : non, L ou L*							
5	Pourcentage = pourcentage des observations							
6	Distance max. = distance maximale au médoïde de la classe							
7	Distance moy. = distance moyenne au médoïde de la classe							
8	Diamètre = distance maximale entre deux observations de la classe							
9	Séparation = distance minimale entre une observation de la classe et une observation d'une autre classe							
10	Silhouette = coefficient moyen de silhouette							
11								
12	Coefficient global de silhouette : 0,87529							
13								
	Classe	Nombre d'observation	Pourcentage	Distance max.	Distance moy.	Diamètre	Séparation	Silhouette
16	Classe 1 (isolée L*)	10	40	0,76699	0,39990	1,45813	5,29886	0,90490
17	Classe 2 (isolée L*)	15	60	1,58977	0,87357	2,62636	5,29886	0,85555
18								
19								
20								
21								

- Classe isolée (non isolée, L ou L*)
- Pourcentage des observations
- Distance maximale au médoïde de la classe
- Distance moyenne au médoïde de la classe
- Diamètre de la classe : distance maximale entre deux observations de la classe
- Séparation de la classe : distance minimale entre une observation de la classe et une observation d'une autre classe
- Coefficient moyen de silhouette

Une classe est une classe L* si son diamètre est plus petit que sa séparation. Une classe est une classe L si pour chaque observation i la dissimilarité maximale entre i et toute autre observation du cluster est plus petite que la dissimilarité minimale entre i et toute observation d'un autre cluster. Une classe L* est donc également une classe L.

Le cinquième tableau affiche pour chaque classe un ensemble de statistiques descriptives pour chacune des variables utilisées pour la classification.

The screenshot shows a window titled 'Rapports et Graphiques' with a tree view on the left. The tree view includes 'Rapport PAM', 'Statistiques pour les classes (PAM)', and 'Statistiques classe 1'. The main area displays a table with 8 columns and 21 rows. The table is titled 'STATISTIQUES POUR LA CLASSE 1' and contains descriptive statistics for two variables, x1 and x2.

	1	2	3	4	5	6	7	8
1								
2	STATISTIQUES POUR LA CLASSE 1							
3								
4		x1	x2					
5	Effectif	10,00000	10,00000					
6	Moyenne	0,08864	0,35503					
7	Variance	0,15401	0,06340					
8	Ecart-type	0,39245	0,25179					
9	MAD (écart absolu moyen)	0,31969	0,20655					
10	Minimum	-0,50750	-0,02759					
11	Maximum	0,85827	0,79960					
12	Etendue	1,36577	0,82319					
13	Médiane	0,04390	0,31919					
14	Premier quartile	-0,26256	0,08682					
15	Troisième quartile	0,28467	0,50482					
16	Dist. inter-quart.	0,54723	0,41799					
17								
18								
19								
20								
21								

Les deux tableaux suivants affichent les contributions signées en pourcentages des variables aux classes et les contributions signées en pourcentages des classes aux variables.

The screenshot shows the 'Rapports et Graphiques' window with the tree view expanded to 'Contrib. variables aux classes (PAM)'. The main area displays a table titled 'CONTRIBUTIONS EN POURCENTAGES DES VARIABLES QUANTITATIVES AUX CLASSES (PAM)'. It includes summary statistics and a table of signed contributions for classes 1 and 2.

	1	2	3	4	5	6	7	8
1								
2	CONTRIBUTIONS EN POURCENTAGES DES VARIABLES QUANTITATIVES AUX CLASSES (PAM)							
3								
4	Variance totale : 274,80087							
5	Variance inter-classes : 264,35247							
6	Pourcentage inter-totale : 96,19783							
7								
8	Les contributions sont signées :							
9	> une valeur négative indique que la variable est inférieure à sa moyenne globale							
10	> une valeur positive indique que la variable est supérieure à sa moyenne globale							
11								
12	Le total non signé en ligne fait 100.							
13								
14								
15		x1	x2					
16	Classe 1	-50,07511	-49,92489					
17	Classe 2	50,07511	49,92489					
18								
19								
20								
21								

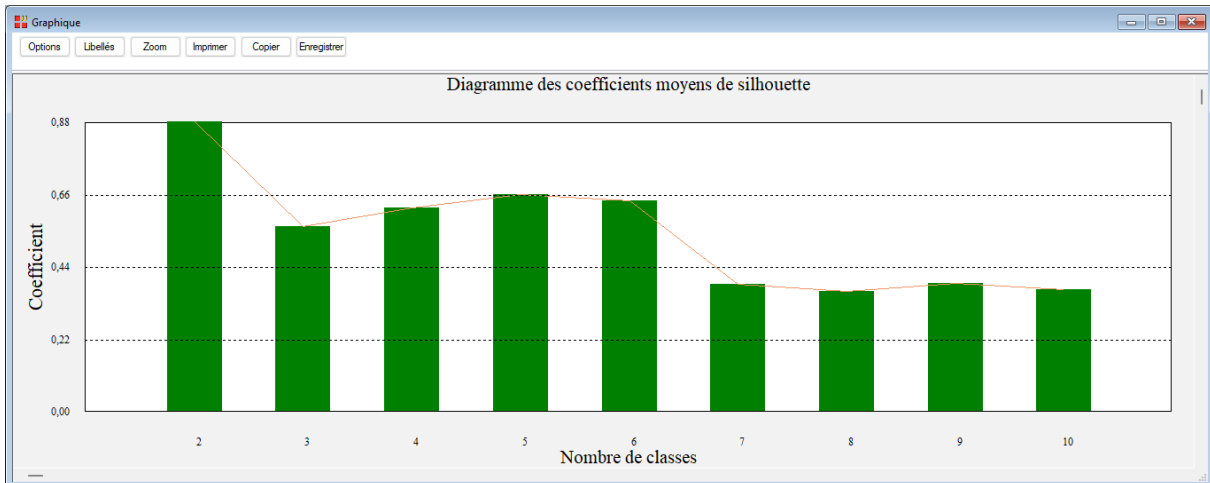
The screenshot shows the 'Rapports et Graphiques' window with the tree view expanded to 'Contrib. classes aux variables (PAM)'. The main area displays a table titled 'CONTRIBUTIONS EN POURCENTAGES DES CLASSES AUX VARIABLES QUANTITATIVES (PAM)'. It includes summary statistics and a table of signed contributions for classes 1 and 2.

	1	2	3	4	5	6	7	8
1								
2	CONTRIBUTIONS EN POURCENTAGES DES CLASSES AUX VARIABLES QUANTITATIVES (PAM)							
3								
4	Les contributions sont signées :							
5	> une valeur négative indique que la variable est inférieure à sa moyenne globale							
6	> une valeur positive indique que la variable est supérieure à sa moyenne globale							
7								
8	Le total non signé en colonne fait 100.							
9								
10								
11		x1	x2					
12	Classe 1	-60	-60					
13	Classe 2	40	40					
14								
15								
16								
17								
18								
19								
20								
21								

L'option Graphiques

- Diagramme des coefficients moyens de silhouette

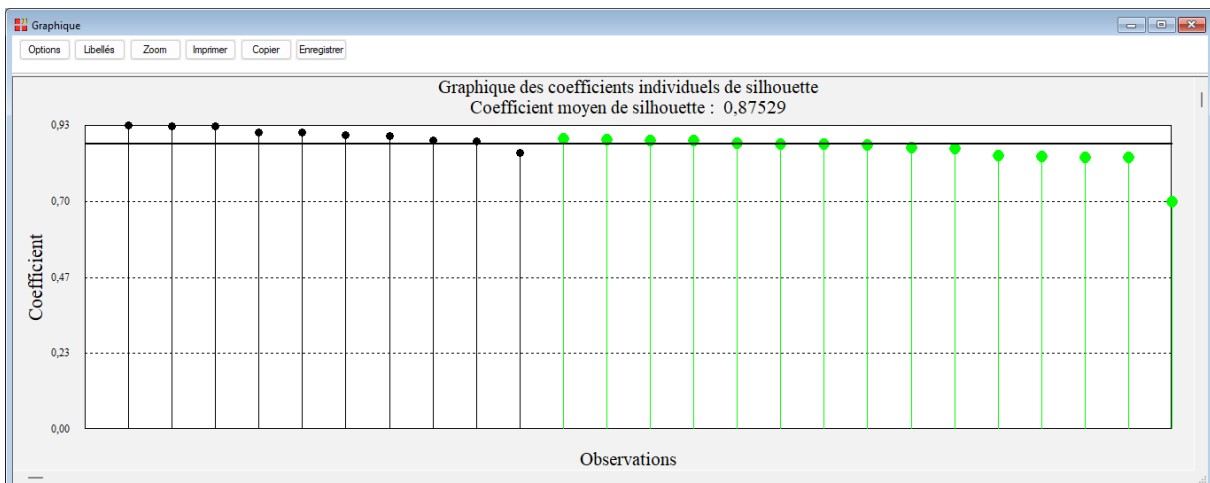
Le diagramme des coefficients moyens de silhouette permet de visualiser l'évolution de ces coefficients en fonction du nombre de classes formées.



- Graphique des coefficients individuels de silhouette

Le graphique des coefficients individuels de silhouette permet de visualiser ces coefficients pour chacune des observations. Pour chaque observation, le coefficient de silhouette est la différence entre la distance moyenne avec les observations de sa classe et la distance moyenne avec les observations des autres classes. Il varie entre -1 et +1. Si négatif, l'observation est en moyenne plus proche de la classe voisine que de la sienne et donc elle est donc mal classée. Si positif, l'observation est en moyenne plus proche de sa classe que de la classe voisine et donc elle est donc bien classée.

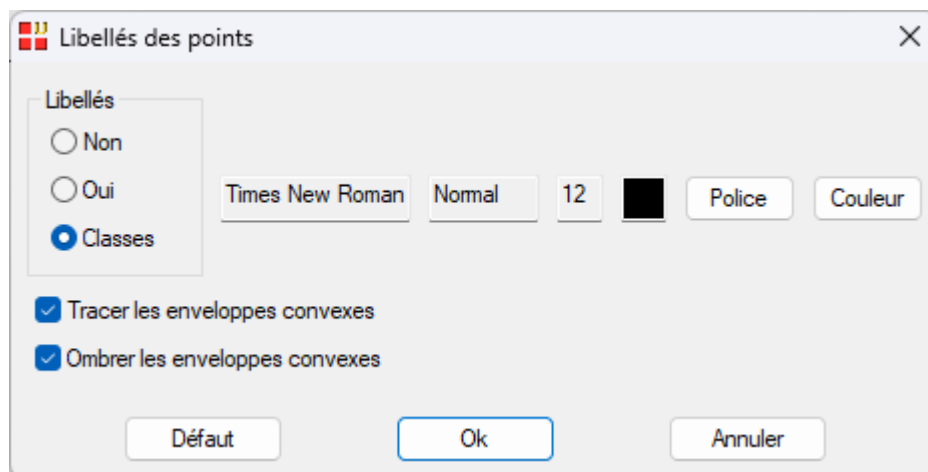
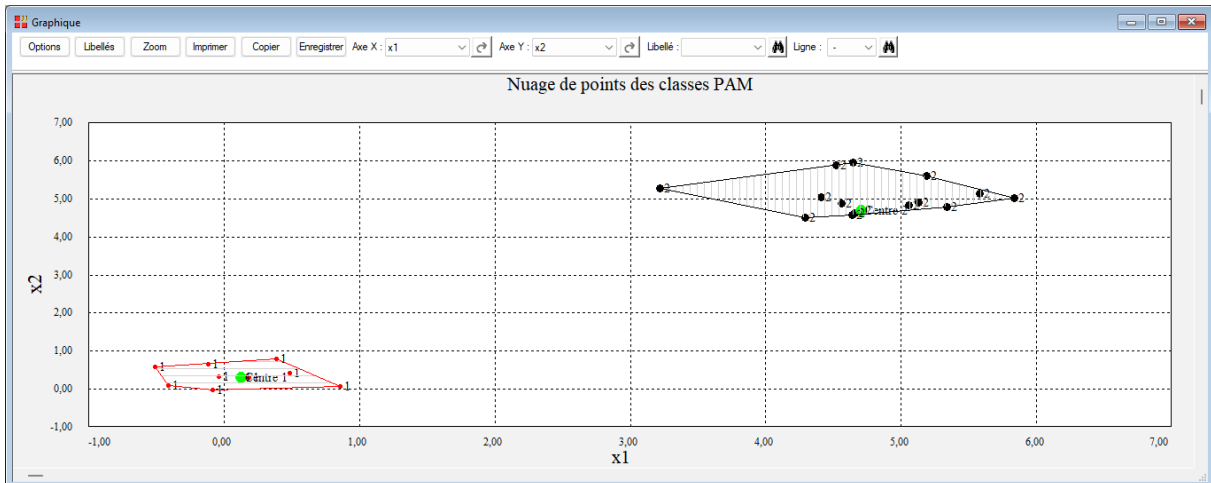
La ligne horizontale indique la position du coefficient moyen de silhouette. Si la plupart des observations d'une classe ont des coefficients inférieurs à ce coefficient moyen, cela indique probablement que la partition obtenue n'est pas bonne.



- Nuage de points des classes

Le graphique du nuage de points des classes permet de visualiser les classes formées par rapport à deux variables quantitatives sélectionnées.

Le bouton 'Libellés' permet de préciser les libellés affichés et si les enveloppes convexes des classes sont tracées ou non.



Exemple 2 : Fichier VEHICULE

Pour illustrer cet exemple, nous utiliserons le fichier VEHICULE.

Ce fichier contient 7 informations caractérisant 24 véhicules : *Modèle, Cylindrée, Puissance, Vitesse, Poids, Longueur, Largeur*.

Les libellés des variables quantitatives sont dans la variable *Mesures*.

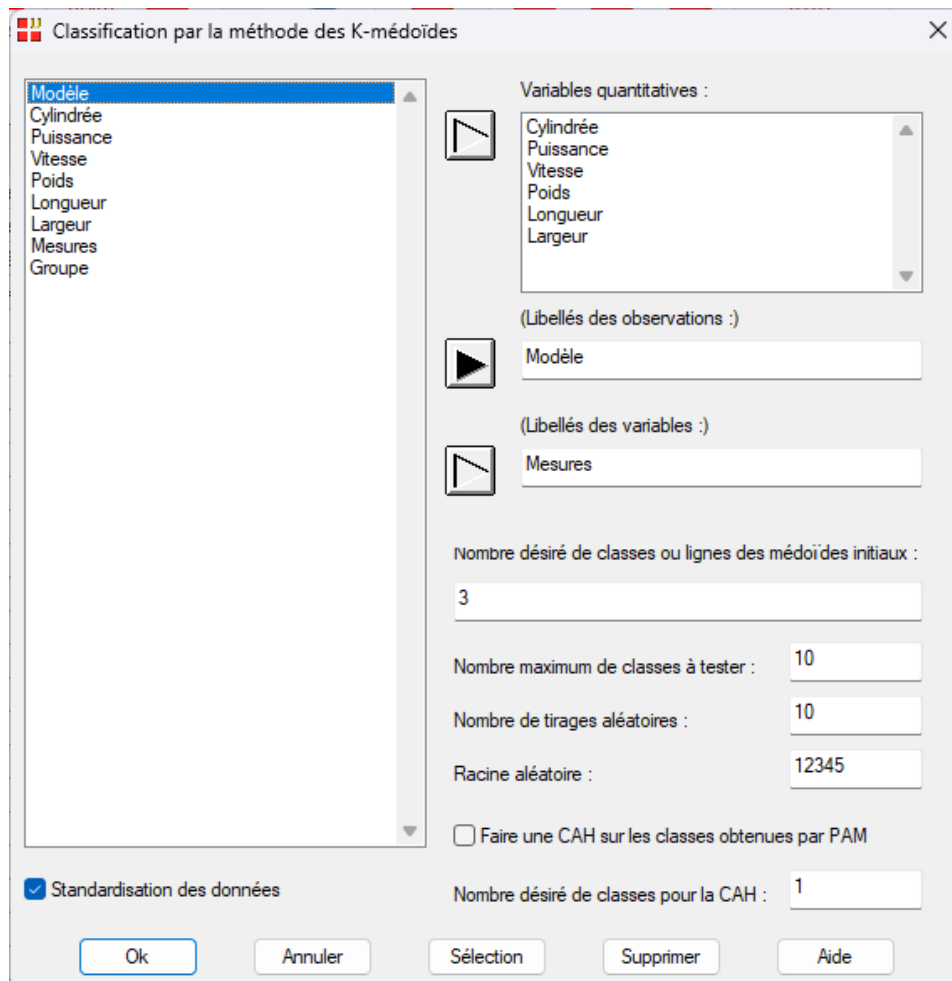
Les 24 modèles d'automobiles sont :

Honda Civic	R19	Fiat Tipo	405
R21	BX	BMW 530i	Rover 827i
R25	Opel Omega	405 Break	Ford Sierra
BMW 325ix	Audi 90 Quattro	Ford Scorpio	Espace
Nissan Vanette	VW Caravelle	Ford Fiesta	Fiat Uno
205	205 Rallye	Seat Ibiza SXI	AX Sport

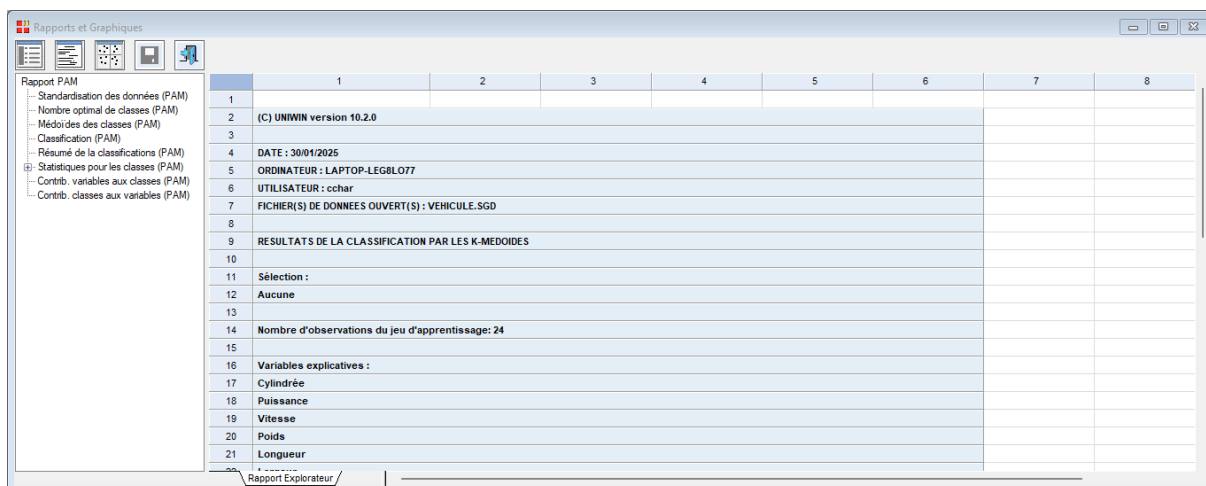
Cliquons sur l'icône KM dans le ruban Décrire et choisissons K-médoïdes. La boîte de dialogue montrée ci-après s'affiche.

Nous choisissons toutes les variables de *Cylindrée* à *Largeur* comme variables quantitatives, la variable *Mesures* comme variable contenant les libellés associés et la variable *Modèle* comme variable contenant les libellés des observations.

Nous demandons de former 3 classes et choisissons de faire une classification sur les données standardisées.



Cliquons sur Ok. UNIWIN débute le calcul de la classification. Après quelques instants, l'écran suivant s'affiche :



Le premier tableau affiche les paramètres de la standardisation des données.

STANDARDISATION DES DONNEES (PAM)		
Moyenne et MAD (écart absolu moyen)		
	Moyenne	MAD
Cylindrée	1906,12500	405,95833
Puissance	113,66667	31,11111
Vitesse	183,08333	18,67361
Poids	1110,83333	197,01389
Longueur	421,58333	35,60417
Largeur	168,83333	5,87500

Le deuxième tableau affiche pour les nombres de classes variant de 2 au nombre maximum de classes indiqué (10), les coefficients moyens de silhouette. Ces coefficients sont également représentés de façon graphique. Ils aident à déterminer le nombre adéquat de classes à former.

DETERMINATION DU NOMBRE OPTIMAL DE CLASSES (PAM)	
Silhouette = coefficient moyen de silhouette	
	Silhouette
2 classes	0,40213
3 classes	0,38557
4 classes	0,34774
5 classes	0,30458
6 classes	0,34514
7 classes	0,32795
8 classes	0,35126
9 classes	0,31775
10 classes	0,27889

Le troisième tableau affiche les médoïdes des classes formées.

MEDOIDES DES CLASSES (PAM)							
	Observation	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Classe 1 - médoïde = Honda-Civic	1	1396	90	174	850	369	166
Classe 2 - médoïde = Ford-Sierra	12	1993	115	185	1190	451	172
Classe 3 - médoïde = Rover-827i	8	2675	177	222	1365	469	175

La classe 1 est représentée par l'observation 1, la classe 2 par l'observation 12 et la classe 3 par l'observation 8.

Le quatrième tableau affiche pour chaque observation sa classe d'affectation, sa classe voisine, sa distance au médoïde et son coefficient de silhouette.

	1	2	3	4	5	6	7	8
1								
2	CLASSIFICATION DES OBSERVATIONS (PAM)							
3								
4	Nombre de classes formées : 3							
5								
6	Distance = distance de l'observation au médoïde de sa classe							
7	Silhouette = coefficient de silhouette de l'observation							
8								
9								
10		Classe affectée	Classe voisine	Distance	Silhouette			
11	Honda-Civic	1	2	0,00000	0,49777			
12	R19	1	2	347,87210	-0,14653			
13	Fiat-Tipo	1	2	221,38880	0,14293			
14	405	2	1	251,10954	0,34765			
15	R21	2	1	97,12363	0,47468			
16	BX	2	1	261,63906	0,19432			
17	BMW-530i	3	2	343,35404	0,60502			
18	Rover-827i	3	2	0,00000	0,62227			
19	R25	3	2	128,15615	0,54319			
20	Opel-Omega	2	3	69,51978	0,33692			
21	405-Break	2	3	113,88591	0,45145			

Le cinquième tableau affiche un résumé de la classification.

	1	2	3	4	5	6	7	8	9
1									
2	INFORMATIONS SUR LES CLASSES (PAM)								
3									
4	Classe isolée : non, L ou L*								
5	Pourcentage = pourcentage des observations								
6	Distance max. = distance maximale au médoïde de la classe								
7	Distance moy. = distance moyenne au médoïde de la classe								
8	Diamètre = distance maximale entre deux observations de la classe								
9	Séparation = distance minimale entre une observation de la classe et une observation d'une autre classe								
10	Silhouette = coefficient moyen de silhouette								
11									
12	Coefficient global de silhouette : 0,3857								
13									
14									
15	Classe	Nombre d'observation	Pourcentage	Distance max.	Distance moy.	Diamètre	Séparation	Silhouette	
16	Classe 1	9	37,50000	2,75628	1,63687	3,94215	0,59579	0,39736	
17	Classe 2	10	41,66667	2,90575	1,48550	4,64854	0,59579	0,32993	
18	Classe 3	5	20,83333	2,34361	1,20933	3,12299	1,63495	0,47565	
19									
20									
21									

- Classe isolée (non isolée, L ou L*)
- Pourcentage des observations
- Distance maximale au médoïde de la classe
- Distance moyenne au médoïde de la classe
- Diamètre de la classe : distance maximale entre deux observations de la classe
- Séparation de la classe : distance minimale entre une observation de la classe et une observation d'une autre classe
- Coefficient moyen de silhouette

Une classe est une classe L* si son diamètre est plus petit que sa séparation. Une classe est une classe L si pour chaque observation i la dissimilarité maximale entre i et toute autre observation du cluster est plus petite que la dissimilarité minimale entre i et toute observation d'un autre cluster. Une classe L* est donc également une classe L.

Le tableau suivant affiche pour chaque classe un ensemble de statistiques descriptives pour chacune des variables utilisées pour la classification.

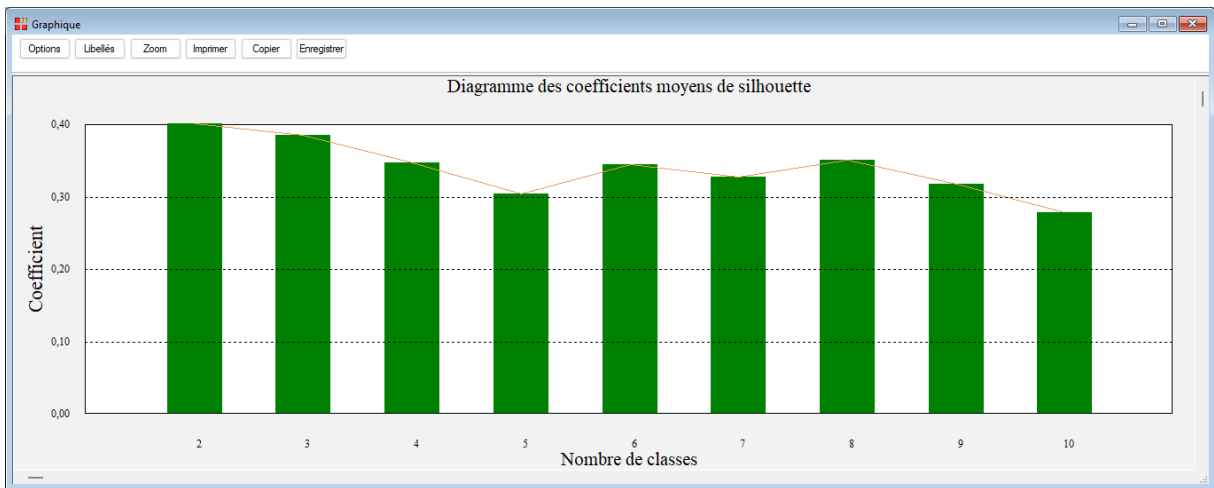
	1	2	3	4	5	6	7	8
1								
2	STATISTIQUES POUR LA CLASSE 1							
3								
4		Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur	
5	Effectif	9,00000	9,00000	9,00000	9,00000	9,00000	9,00000	9,00000
6	Moyenne	1395,44444	83,44444	168,55556	857,22222	374,11111	161,77778	
7	Variance	39400,91358	298,24691	305,13580	6289,50617	331,65432	27,06173	
8	Ecart-type	198,49663	17,28963	17,46814	79,30841	18,21138	5,20209	
9	MAD (écart absolu moyen)	169,06173	13,95062	14,81481	69,13580	13,72840	4,41975	
10	Minimum	1116,00000	50,00000	135,00000	730,00000	350,00000	155,00000	
11	Maximum	1721,00000	103,00000	189,00000	970,00000	415,00000	170,00000	
12	Etendue	605,00000	53,00000	54,00000	240,00000	65,00000	15,00000	
13	Médiane	1345,00000	86,50000	172,00000	830,00000	369,50000	160,50000	
14	Premier quartile	1205,50000	69,00000	152,00000	792,50000	363,50000	156,50000	
15	Troisième quartile	1520,50000	93,50000	180,50000	902,50000	370,50000	164,00000	
16	Dist. inter-quart.	315,00000	24,50000	28,50000	110,00000	7,00000	7,50000	

Les deux tableaux suivants affichent les contributions signées en pourcentages des variables aux classes et les contributions signées en pourcentages des classes aux variables.

	1	2	3	4	5	6	7	8
1								
2	CONTRIBUTIONS EN POURCENTAGES DES VARIABLES QUANTITATIVES AUX CLASSES (PAM)							
3								
4	Variance totale : 7719474,29167							
5	Variance inter-classes : 6828729,62500							
6	Pourcentage inter/totale : 88,46107							
7								
8	Les contributions sont signées :							
9	> une valeur négative indique que la variable est inférieure à sa moyenne globale							
10	> une valeur positive indique que la variable est supérieure à sa moyenne globale							
11								
12	Le total non signé en ligne fait 100.							
13								
14								
15		Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur	
16	Classe 1	-79,37962	-0,27801	-0,06424	-19,57703	-0,68595	-0,01515	
17	Classe 2	19,58891	-0,06226	-0,10444	76,00519	4,12380	0,11540	
18	Classe 3	89,91456	0,47907	0,14804	9,23690	0,21786	0,00356	

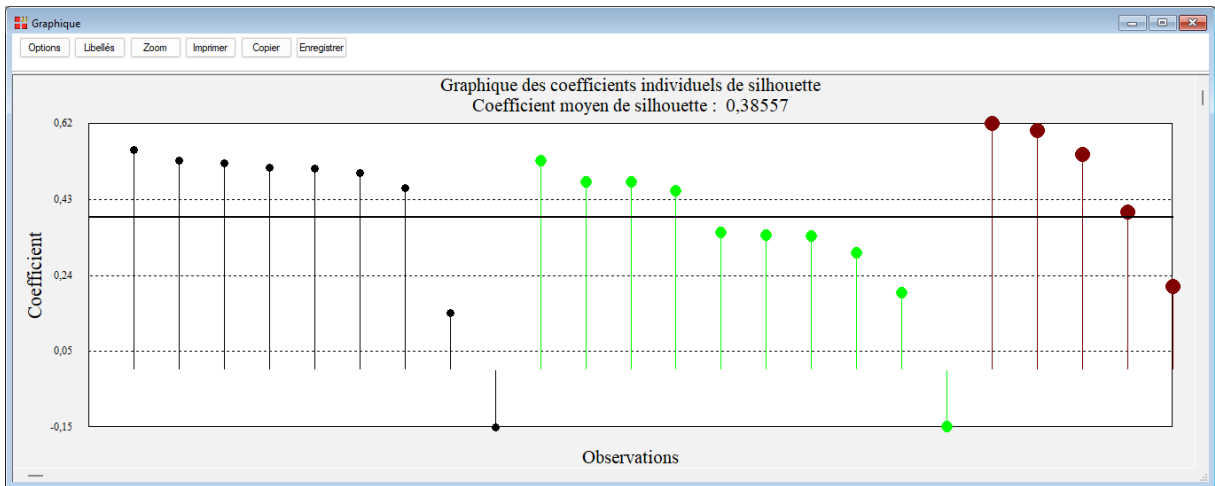
	1	2	3	4	5	6	7	8
1								
2	CONTRIBUTIONS EN POURCENTAGES DES CLASSES AUX VARIABLES QUANTITATIVES (PAM)							
3								
4	Les contributions sont signées :							
5	> une valeur négative indique que la variable est inférieure à sa moyenne globale							
6	> une valeur positive indique que la variable est supérieure à sa moyenne globale							
7								
8	Le total non signé en colonne fait 100.							
9								
10								
11		Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur	
12	Classe 1	-40,87650	-31,30762	-25,06637	-56,82997	-60,50859	-61,93552	
13	Classe 2	0,41942	-0,29152	-1,69443	9,17385	15,12529	19,61323	
14	Classe 3	58,70407	68,40086	73,23920	33,99618	24,36612	18,45125	
15								
16								
17								
18								
19								
20								
21								

Le diagramme des coefficients moyens de silhouette permet de visualiser l'évolution de ces coefficients en fonction du nombre de classes formées.



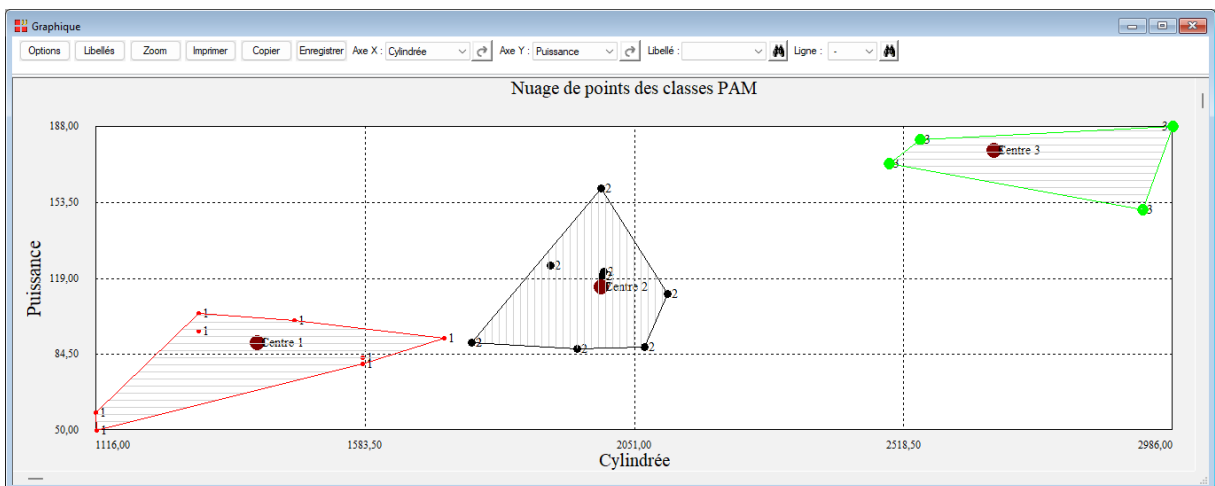
Le graphique des coefficients individuels de silhouette permet de visualiser ces coefficients pour chacune des observations. Pour chaque observation, le coefficient de silhouette est la différence entre la distance moyenne avec les observations de sa classe et la distance moyenne avec les observations des autres classes. Il varie entre -1 et +1. Si négatif, l'observation est en moyenne plus proche de la classe voisine que de la sienne et donc elle est donc mal classée. Si positif, l'observation est en moyenne plus proche de sa classe que de la classe voisine et donc elle est donc bien classée.

La ligne horizontale indique la position du coefficient moyen de silhouette. Si la plupart des observations d'une classe ont des coefficients inférieurs à ce coefficient moyen, cela indique probablement que la partition obtenue n'est pas bonne.



Le graphique du nuage de points des classes permet de visualiser les classes formées par rapport à deux variables quantitatives sélectionnées.

Le bouton 'Libellés' permet de préciser les libellés affichés et si les enveloppes convexes des classes sont tracées ou non.



Libellés des points [X]

Libellés

Non

Oui

Classes

Times New Roman Normal 12 [Couleur] [Police] [Couleur]

Tracer les enveloppes convexes

Ombre les enveloppes convexes

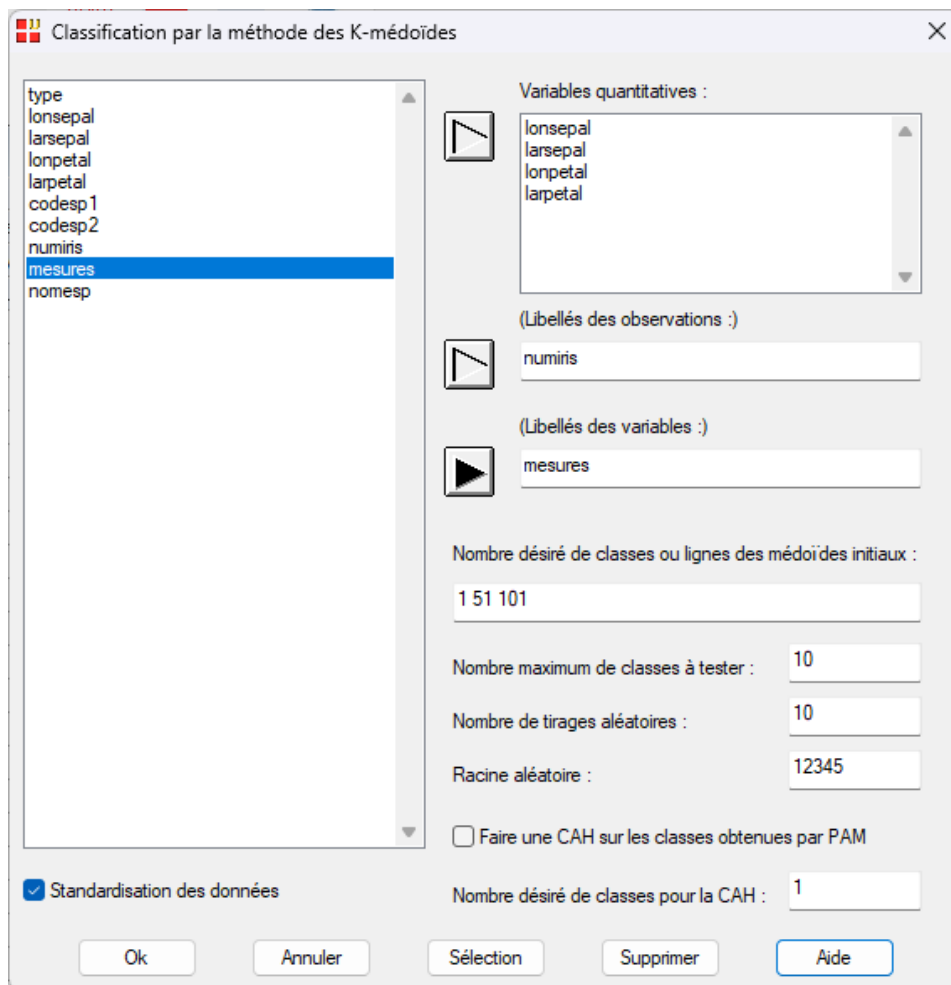
[Défaut] [Ok] [Annuler]

Exemple 3 : Fichier IRIS3

Nous utiliserons le fichier IRIS3 pour illustrer ce troisième exemple.

Ce fichier contient pour 150 iris les mesures des quatre caractéristiques suivantes exprimées en millimètres : longueur du sépale, largeur du sépale, longueur du pétale et largeur du pétale

Renseignons la boîte de dialogue comme montré ci-dessous en précisant les médoïdes initiaux à utiliser : les lignes 1, 51 et 101 du fichier des données.

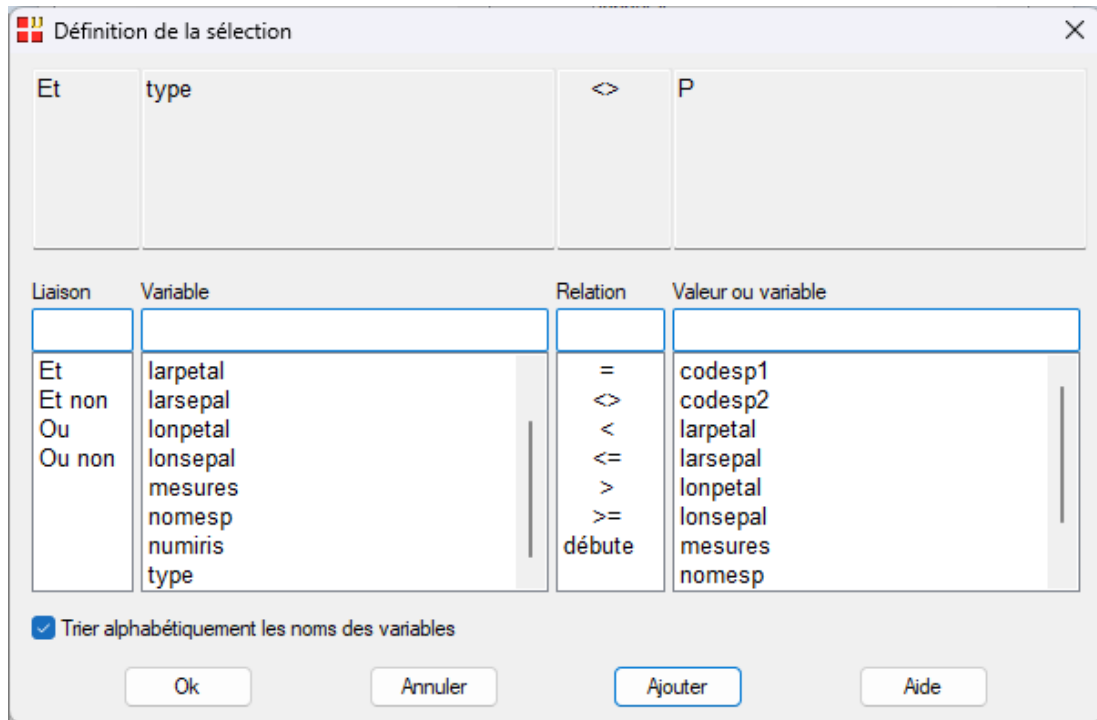


Cliquons sur le bouton 'Sélection' pour sélectionner les observations à utiliser comme jeu d'apprentissage.

Un message nous indique que 144 observations seront utilisées.

Cliquons sur Ok.

Un message nous indique que les lignes non sélectionnées seront utilisées comme jeu de prévision.



Visualisons quelques résultats obtenus.

	1	2	3	4	5	6	7	8
1								
2	DETERMINATION DU NOMBRE OPTIMAL DE CLASSES (PAM)							
3								
4	Silhouette = coefficient moyen de silhouette							
5								
6								
7								
8	2 classes		Silhouette					
9	3 classes		0,56713					
10	4 classes		0,45212					
11	5 classes		0,40200					
12	6 classes		0,34280					
13	6 classes		0,34497					
14	7 classes		0,33317					
15	8 classes		0,31047					
16	9 classes		0,33572					
17	10 classes		0,35812					
18								
19								
20								
21								

	1	2	3	4	5	6	7	8
1								
2	MÉDOIDES DES CLASSES (PAM)							
3								
4								
5								
6		Observation	lonsepal	larsepal	lonpetal	larpetal		
7	Classe 1 - médoïde = 8	7	5,0	3,4	1,5	0,2		
8	Classe 2 - médoïde = 95	91	5,6	2,7	4,2	1,3		
9	Classe 3 - médoïde = 148	142	6,5	3,0	5,2	2,0		
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
- Contrib. variables aux classes (PAM)
- Contrib. classes aux variables (PAM)
- Classification du jeu de prévision (PAM)

	1	2	3	4	5	6	7	8
1								
2	CLASSIFICATION DES OBSERVATIONS (PAM)							
3								
4	Nombre de classes formées : 3							
5								
6	Distance = distance de l'observation au médoïde de sa classe							
7	Silhouette = coefficient de silhouette de l'observation							
8								
9								
10		Classe affectée	Classe voisine	Distance	Silhouette			
11	1	1	2	0,17321	0,73278			
12	2	1	2	0,42426	0,50289			
13	4	1	2	0,50000	0,57300			
14	5	1	2	0,22361	0,72989			
15	6	1	2	0,70000	0,62523			
16	7	1	2	0,42426	0,68509			
17	8	1	2	0,00000	0,72179			
18	9	1	2	0,78740	0,40673			
19	10	1	2	0,33166	0,58493			
20	11	1	2	0,50000	0,68306			
21	12	1	2	0,22361	0,71105			

Rapport Explorateur /

Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
- Contrib. variables aux classes (PAM)
- Contrib. classes aux variables (PAM)
- Classification du jeu de prévision (PAM)

	1	2	3	4	5	6	7	8	9
1									
2	INFORMATIONS SUR LES CLASSES (PAM)								
3									
4	Classe isolée : non, L ou L*								
5	Pourcentage = pourcentage des observations								
6	Distance max. = distance maximale au médoïde de la classe								
7	Distance moy. = distance moyenne au médoïde de la classe								
8	Diamètre = distance maximale entre deux observations de la classe								
9	Séparation = distance minimale entre une observation de la classe et une observation d'une autre classe								
10	Silhouette = coefficient moyen de silhouette								
11									
12	Coefficient global de silhouette : 0,45212								
13									
14									
15	Classe	Nombre d'observation	Pourcentage	Distance max.	Distance moy.	Diamètre	Séparation	Silhouette	
16	Classe 1	47	32,63889	3,09924	0,97998	4,76725	1,76193	0,63085	
17	Classe 2	41	28,47222	3,11136	1,04642	4,09799	0,28899	0,39827	
18	Classe 3	56	38,88889	3,18131	1,16102	4,53052	0,28899	0,34155	
19									
20									
21									

Rapport Explorateur /

Un tableau indique les classes affectées aux données du jeu de prévision :

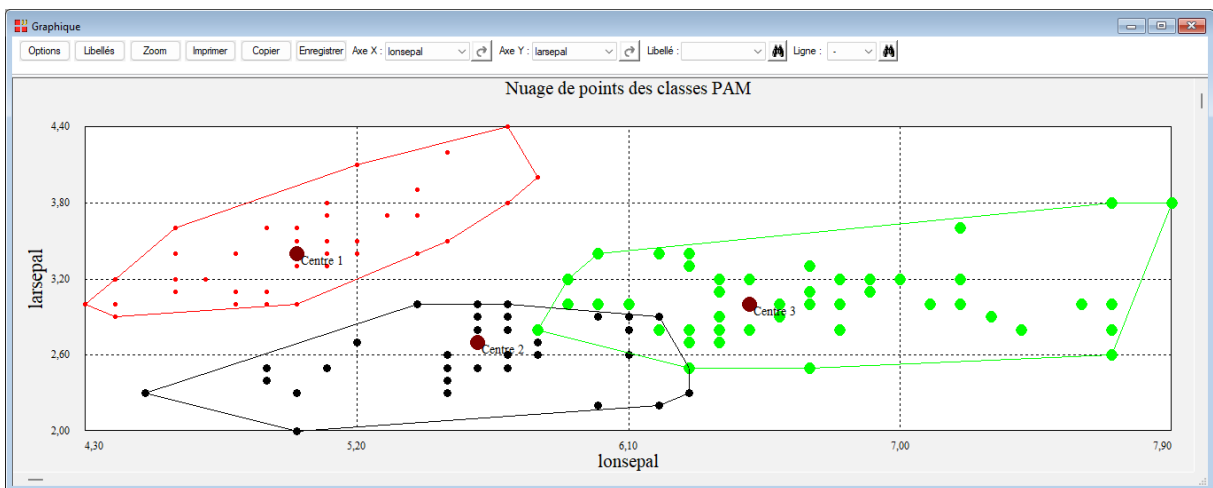
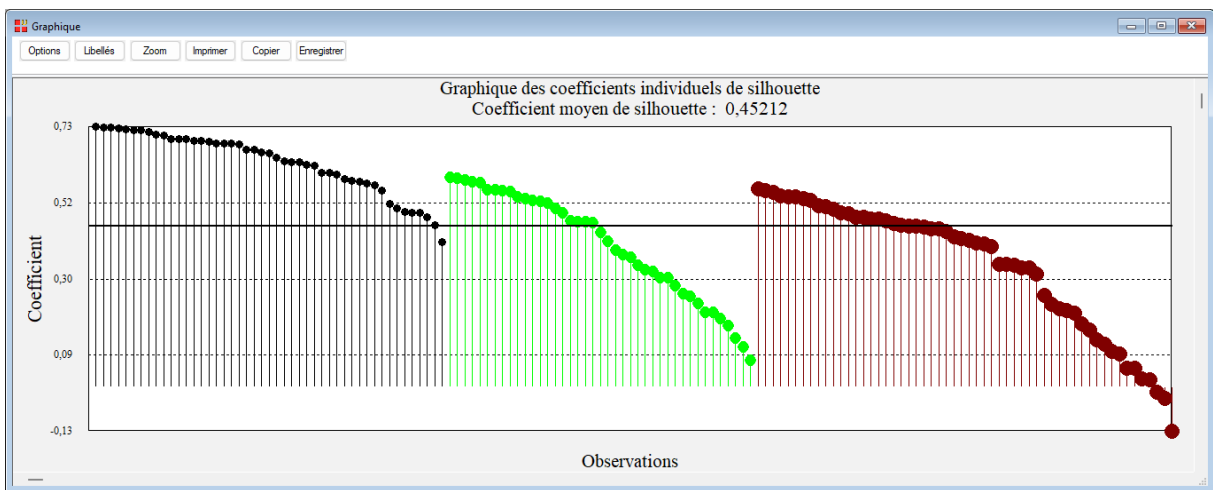
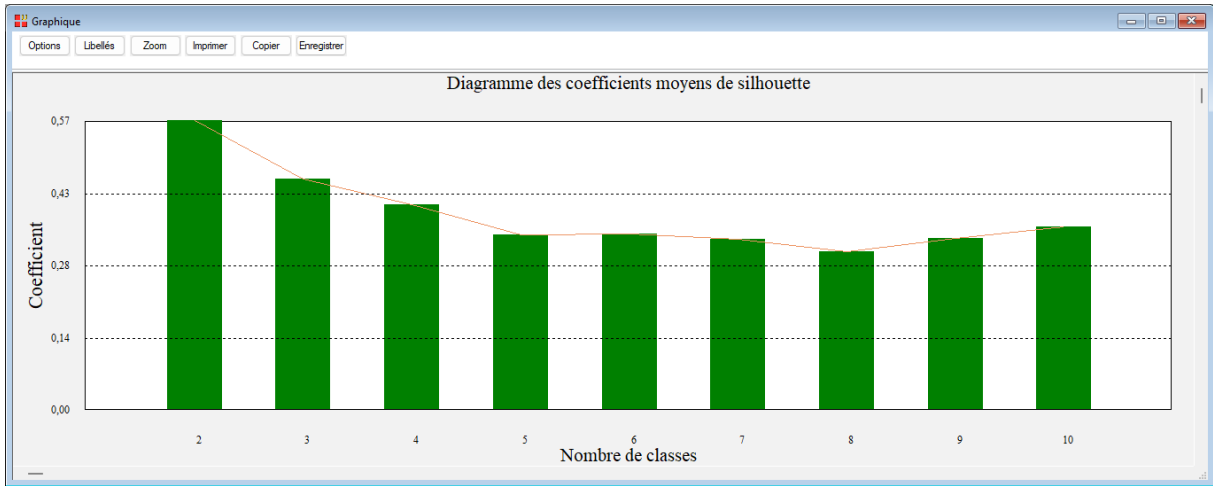
Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
- Contrib. variables aux classes (PAM)
- Contrib. classes aux variables (PAM)
- Classification du jeu de prévision (PAM)

	1	2	3	4	5	6	7	8
1								
2	CLASSIFICATION DES OBSERVATIONS DU JEU DE PREVISION (PAM)							
3								
4	Classe affectée et distances de l'observation aux médoïdes des classes							
5								
6								
7		Classe affectée	Classe 1	Classe 2	Classe 3			
8	3	1	0,41231	3,26803	4,66154			
9	36	1	0,36056	3,28938	4,64004			
10	62	2	3,15436	0,46904	1,26888			
11	84	3	4,05093	1,02956	0,71414			
12	104	3	4,61828	1,65529	0,50000			
13	125	3	4,91426	2,11187	0,62450			
14								
15								
16								
17								
18								
19								
20								
21								

Rapport Explorateur /

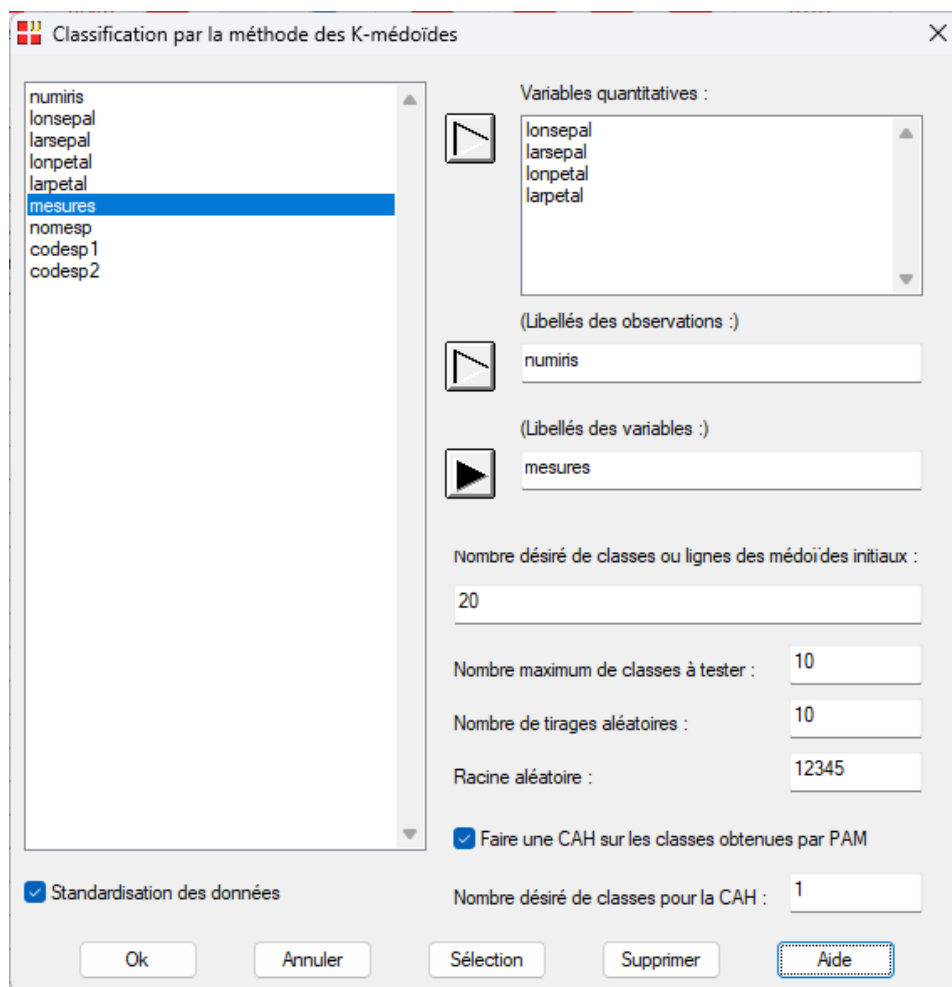


Exemple 4 : Fichier IRIS - Classification mixte

Si le nombre d'observations à classer est important, la classification mixte est une démarche intéressante. Elle se déroule en trois étapes :

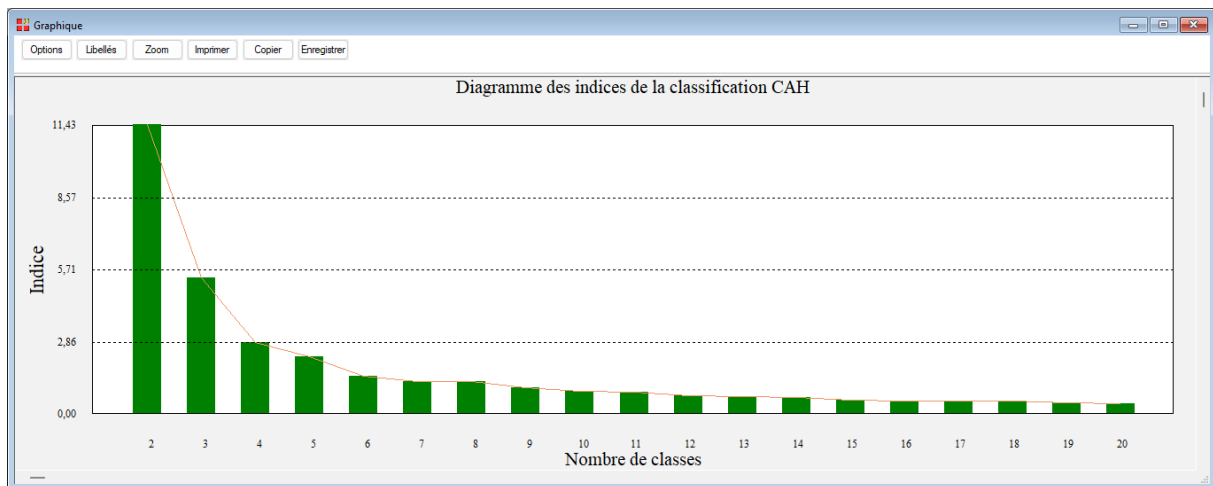
1. Partitionnement en q classes par la méthode K-médoïdes
2. Classification ascendante hiérarchique (CAH) sur les q classes issues des K-médoïdes avec pondération par les effectifs des classes
3. Partition finale obtenue par troncature de l'arbre CAH

A titre d'exemple illustratif, mettons cette démarche en œuvre en utilisant le fichier IRIS et en demandant un nombre de classes égal à 20 (le nombre maximum de classes à tester est alors automatiquement défini à 20).



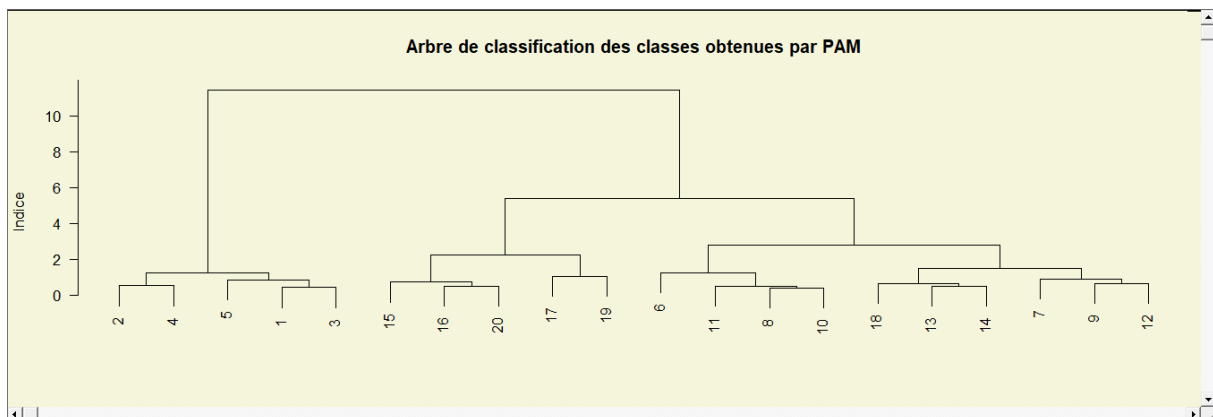
Cochons la case 'Faire une classification CAH sur les classes obtenues par PAM' et demandons 1 classe pour la CAH.


Affichons le diagramme des indices de la classification CAH.



Ce graphique nous indique que 3 classes est un choix possible.

Visualisons l'arbre de la classification.



Revenons à la boîte de dialogue d'entrée des données en cliquant sur  et indiquons un nombre de classes égal à 3 pour la classification CAH.

Faire une CAH sur les classes obtenues par PAM

Nombre désiré de classes pour la CAH :

Exécutons à nouveau l'analyse et visualisons les résultats obtenus dans le rapport ainsi que le graphique du nuage des points des classes de la CAH.

Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
- Contrib. variables aux classes (PAM)
- Contrib. classes aux variables (PAM)
- Indices de la classification (CAH)**
- Centroides des classes (CAH)
- Classification (CAH)
- Résumé de la classification (CAH)
- Statistiques pour les classes (CAH)
- Contrib. variables aux classes (CAH)
- Contrib. classes aux variables (CAH)

	1	2	3	4	5	6	7	8
1								
2	INDICES DE LA CLASSIFICATION (CAH)							
3								
4	La classification ascendante hiérarchique a été faite sur les médoïdes							
5	des classes formées par la classification par les K-médoïdes.							
6								
7								
8			Indice	Variation				
9	2 classes		11,42516					
10	3 classes		5,36910	6,05606				
11	4 classes		2,80395	2,56516				
12	5 classes		2,23046	0,57348				
13	6 classes		1,48748	0,74298				
14	7 classes		1,25706	0,23042				
15	8 classes		1,25488	0,00218				
16	9 classes		1,01980	0,23507				
17	10 classes		0,88765	0,13216				
18	11 classes		0,83811	0,04954				
19	12 classes		0,70993	0,12818				
20	13 classes		0,65502	0,05491				
21	14 classes		0,61644	0,03858				

Rapport Explorateur

Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
- Contrib. variables aux classes (PAM)
- Contrib. classes aux variables (PAM)
- Centroides des classes (CAH)**
- Classification (CAH)
- Résumé de la classification (CAH)
- Statistiques pour les classes (CAH)
- Contrib. variables aux classes (CAH)
- Contrib. classes aux variables (CAH)

	1	2	3	4	5	6	7	8
1								
2	CENTROIDES DES CLASSES (CAH)							
3								
4								
5			lonsepal	larsepal	lonpetal	larpetal		
6	Classe 1		5,01633	3,45102	1,46531	0,24490		
7	Classe 2		5,55417	2,75278	4,48056	1,46369		
8	Classe 3		6,96552	3,14828	5,83793	2,15517		
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Rapport Explorateur

Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
- Contrib. variables aux classes (PAM)
- Contrib. classes aux variables (PAM)
- Indices de la classification (CAH)
- Centroides des classes (CAH)
- Classification (CAH)**
- Résumé de la classification (CAH)
- Statistiques pour les classes (CAH)
- Contrib. variables aux classes (CAH)
- Contrib. classes aux variables (CAH)

	1	2	3	4	5	6	7	8
1								
2	CLASSIFICATION DES OBSERVATIONS (CAH)							
3								
4	Nombre de classes CAH : 3							
5	Nombre de classes PAM : 20							
6								
7								
8			Classe CAH	Classe PAM				
9	1		1	1				
10	2		1	2				
11	3		1	2				
12	4		1	2				
13	5		1	1				
14	6		1	3				
15	7		1	1				
16	8		1	1				
17	9		1	4				
18	10		1	2				
19	11		1	3				
20	12		1	1				
21	13		1	2				

Rapport Explorateur

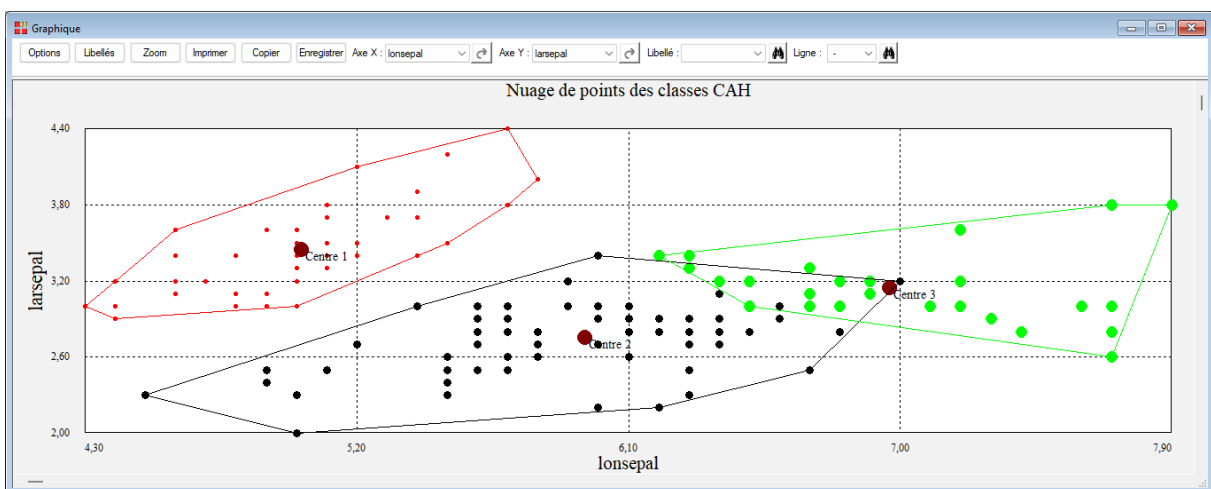
Rapports et Graphiques

Rapport PAM

- Standardisation des données (PAM)
- Nombre optimal de classes (PAM)
- Médoïdes des classes (PAM)
- Classification (PAM)
- Résumé de la classifications (PAM)
- Statistiques pour les classes (PAM)
 - Contrib. variables aux classes (PAM)
 - Contrib. classes aux variables (PAM)
- Indices de la classification (CAH)
- Centroides des classes (CAH)
- Classification (CAH)
- Résumé de la classification (CAH)**
- Statistiques pour les classes (CAH)
 - Contrib. variables aux classes (CAH)
 - Contrib. classes aux variables (CAH)

	1	2	3	4	5	6	7	8
1								
2	RESUME DE LA CLASSIFICATION (CAH)							
3								
4								
5		Nombre d'observations	Pourcentage					
6	Classe 1	49	32,66667					
7	Classe 2	72	48,00000					
8	Classe 3	29	19,33333					
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Rapport Explorateur /



Il est possible d'enregistrer ces résultats :

Enregistrement des résultats (1/1)

Enregistrer

- Libellés des observations
- Libellés des variables
- Classes des observations (PAM)
- Libellés des classes (PAM)
- Médoïdes des classes (PAM)
- Indices de la classification (CAH)
- Classes des observations (CAH)
- Centroides des classes (CAH)

Noms attribués aux variables cibles

libobs

libvar

classes

libclasses

medoides_1

indicescah

classescah

centrescah_1

Ok Plus Tout Annuler

Les variables internes créées par la procédure

Voici la liste des variables internes créées par la procédure.

<i>Variable</i>	<i>Contenu</i>
libobs	Libellés des observations du jeu d'apprentissage
libvar	Libellés des variables
classes	Classes des observations du jeu d'apprentissage (PAM)
libclasses	Libellés des classes (PAM)
medoides	Médoïdes des classes (PAM)
libobsprev	Libellés des observations du jeu de prévision (PAM)
classesprev	Classes affectées aux observations du jeu de prévision (PAM)
indicescah	Indices de la classification CAH (classification mixte)
classescah	Classes des observations CAH (classification mixte)
centrescah	Centroïdes non standardisés des classes CAH (classification mixte)

Références

Documentation du package R 'stats' (2024)

<https://rdr.io/r/stats/stats-package.html>

Documentation du package R 'cluster' (2024)

<https://cran.r-project.org/web/packages/cluster/cluster.pdf>